画像/言語同時埋め込みベクトル空間の構築に向けた埋め込み粒度の比較検討

北山 晃太郎 [1]　清野 舜 [1,2]　鈴木 潤 [1,2]　乾 健太郎 [1,2]
[1] 東北大学　[2] 理化学研究所 AIP センター
{kitayama,kiyono,jun.suzuki,inui}@ecei.tohoku.ac.jp

1 はじめに

現在，言語，画像，音声といった情報処理関連の研究分野では，深層ニューラルネットワークに基づく方法論が主流となっており，多くの研究成果が報告されている．歴史的には，各分野の課題に特化した技術が研究開発されてきたが，深層ニューラルネットワークの技術発展に伴って，各分野で用いられる技術がほぼ同一のものとなった．こういった背景により，複数分野を組み合わせた研究への取り組みが容易になり，多くのマルチモーダルタスクが盛んに研究されるようになってきた．例えば，言語と画像を組み合わせたマルチモーダルタスクとして，画像キャプション生成 [1] や画像–文検索 [2] などが挙げられる．

近年，多くの訓練済みニューラルモデルが再利用可能なコンポーネントとして配布されている．これらは，様々なタスクに汎用的に適用可能であり，研究分野の発展に大きな貢献をもたらしてきた．例えば，言語処理研究分野においては，訓練済みニューラル言語モデル及びその派生技術が大きな役割を担ってきた．代表的なものとして，Common Crawl データから訓練された単語埋め込みベクトル（GloVe）[3] がある．また，最近では ELMo[4], BERT[5] といった発展的な訓練済み言語モデルが公開され，様々な自然言語処理タスクの性能を大幅に向上できることを示している．同様に，画像処理の分野でも，VGG16[6] や ResNet152[7] といった画像特徴抽出用の訓練済みモデルが配布されており，画像を扱うタスクの基盤的な資源として広く用いられている．

本研究の目的は，画像–言語のマルチモーダルタスクにおいて，前述の GloVe や VGG16 のような訓練済みモデルに相当する再利用可能なコンポーネントを構築することである．本稿では，その一例として，Visual Word2Vec[8] のように画像と言語の情報を共通のベクトル空間（以下，共通空間と呼ぶ）へ埋め込むことを考える．

表1: 共通空間への埋め込みにおける入力の粒度の比較

	単語	説明文
画像全体	–	Kiros[9]
物体画像	Frome[10]	Karpathy[11, 12]

2 関連研究

言語と画像の共通空間への埋め込みを考えるにあたり，各情報をどのような “粒度” で利用すればよいかを考慮する必要がある．例えば，共通空間を用いて画像の間違い探しタスク [13] を解くことを考える．このとき，画像内の物体を捉えることが重要だと考えられるため，直感的には物体（オブジェクト）に区切られた画像が空間上で利用可能であることが望ましい．

本節では，関連研究を，埋め込む情報の粒度に着目して分類し，表1にまとめた．言語の観点では，埋め込みに用いる情報として，単語を用いるか説明文（フレーズを含む）を用いるか，という粒度で分類した．一方，画像の観点では，画像全体を用いる場合と，特定の物体のみが表示されている画像（以降，物体画像と呼ぶ）を用いる場合で分類した．

以下，分類した結果を元にして関連研究について述べる．

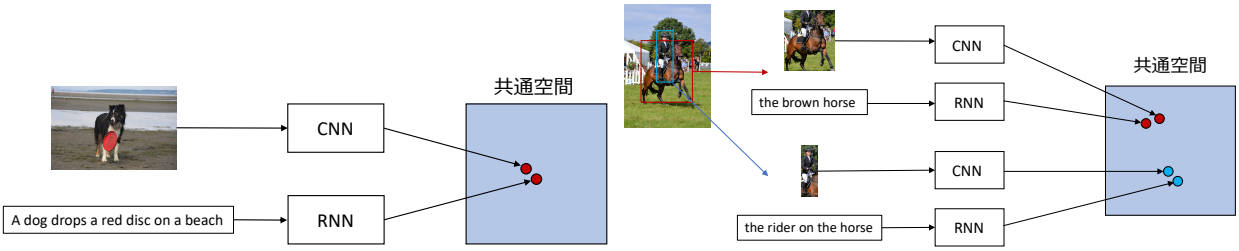
2.1 物体画像と単語の共通空間埋め込み

物体画像と単語を共通空間に埋め込む手法として，Frome ら [10] のモデル（DeViSE）が挙げられる．埋め込みの訓練時には，画像と単語の組がベクトル空間上で近くに配置されるような最適化が行われる．画像とラベルの特徴抽出には，画像側は ImageNet で訓練した CNN，ラベル側は skip-gram 言語モデルで訓練した単語ベクトルが用いられている．同手法は，ImageNet object recognition challenge[*1]における当時の最高性能を達成した．

2.2 画像全体と説明文の共通空間埋め込み

Kiros ら [9] は，一般画像と説明文を対象として，共通空間への埋め込みを行っている．画像と説明文の特

[*1] http://www.image-net.org/challenges/LSVRC/



(a) 画像全体をそのまま埋め込む場合　　**(b)** 画像を物体画像に区切って埋め込む場合

図1: 画像と文を共通空間へ埋め込む際の，情報の粒度の違い

表2: 各データセットに含まれる画像の枚数の比較

	画像の粒度	訓練	開発	評価
Flickr8K	全体画像	6,091	1,000	1,000
Flickr30K	全体画像	25,783	3,000	3,000
Flickr30K Entities	物体画像	247,599	32,217	32,247
Visual Genome	物体画像	3,034,017	473,801	15,000

徴抽出には，それぞれ ImageNet で訓練した CNN と，LSTM を用いている．また，訓練および評価データには，Flickr8K[14] と Flickr30K[15] を用いている．

2.3 物体画像と説明文の共通空間埋め込み

Karpathy ら [11, 12] は，物体画像と説明文を埋め込む方法を提案している．画像は，物体検出のネットワークを用いて物体画像への切り出しを行った後，訓練済みの CNN を用いて特徴抽出を行う．言語側は，説明文を依存構造解析器に入力し，得られた係り受け関係をタプルの形に変換した後に，埋め込みを行う．

実験では，Flickr8K と Flickr30K に加えて，Pascal1K[16] や MSCOCO[17] を用いている．

3 実験

本実験では，画像と言語を共通空間に埋め込む場合に，各情報の粒度がベクトル空間に及ぼす影響を検証する．まず，画像と言語について (a) 画像全体と説明文（図1a）(b) 物体画像と説明文（図1b）という 2 つの粒度についての埋め込みを訓練する．その後，各空間を定量的・定性的に比較することで，その特徴を明らかにする．

3.1 データセット

今回用いたデータセットと，その中に含まれる画像の枚数を表2にまとめた．

一般画像と説明文を共通空間に埋め込むためのデータセットとしては，Flickr8K と Flickr30K を用いた．両データセットには，画像 1 枚に対して，対応する説明文が 5 文付与されている．今回は画像と各説明文とのペアを独立のインスタンスとみなし，同時埋め込みの訓練に用いた．また，既存研究 [9, 11, 12] に従って訓練・開発・評価データへの分割を行った．

また，画像内の物体ごとに共通空間に埋め込むためのデータセットとしては Visual Genome[*2][18] と Flickr30K Entities[19] を用いた．これらのデータセットでは，画像中に物体の矩形情報とその説明文が付与されている．矩形情報から画像を抽出することで，物体画像と説明文の対応のついたデータセットとして利用した．Flickr30K Entities に関しては，Flickr30K と同じように訓練・開発・評価データへ分割を行った．また，Visual Genome に関しては，データセット中に含まれる 91,039 枚の画像[*3]を表2に示した通りにランダムに分割した．その後，評価データとしては，10,000 枚から獲得した物体画像と説明文全 359,489 組のうち，ランダムに 15,000 組を抽出した．

3.2 実験設定

画像と文の共通空間への埋め込みを訓練するにあたり，Kiros ら [9] のモデルを用いた[*4]．全ての実験に同じハイパーパラメータを用いた．具体的には，エポック数 100，バッチサイズ 1,024，埋め込み先の次元数を 1,000 と設定した．最適化手法には Adam を採用し，α の値は 0.0002 とした．また，画像からの特徴抽出には，訓練済みの VGG16[6] を適用し，4096 次元の特徴ベクトルを獲得した．

4 実験結果

4.1 定量評価：画像–文検索

定量評価として，訓練した共通空間を用いて (a) 文から画像の検索タスク (b) 画像から文の検索タスクに取り組んだ．検索結果のランキングは，共通空間上で入力ベクトルと他のデータ点とのコサイン距離を計算し，距離が近いものから順にソートすることによって求めた．

評価指標としては，先行研究 [9, 11, 12] と同様に Recall@K と Med r を用いた．ここで Recall@K は，ランク付けの際に，正解のものがランキングの上位 K 個以内に入っている割合を表すもので，値が高いほど高い

[*2] https://visualgenome.org/

[*3] Visual Genome には合計 108,077 枚の画像が含まれるが，そのうち矩形情報が付与されているのが 91,039 枚であった．

[*4] モデルの実装として https://github.com/josharnoldjosh/Image-Caption-Joint-Embedding を用いた．

表3: 各モデルの Recall@K（R@K）と Median rank （Med r）の比較：Entities は Flickr30K Entities を表す

評価データ：Flickr30K

	文検索				画像検索			
訓練データ	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr8K	6.3	20.3	30.1	31.0	3.6	9.9	14.2	79.2
Flickr30K	11.7	33.9	47.5	11.5	6.1	14.4	19.0	77.0
Flickr30K(+Entities)	11.8	33.7	46.9	12.1	5.4	13.4	18.2	78.0
Entities	1.9	6.9	12.3	78.9	1.2	4.7	8.5	86.7
Visual Genome	3.9	12.7	20.3	49.5	2.6	8.2	13.7	81.5

評価データ：Visual Genome

	文検索				画像検索			
訓練データ	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Flickr8K	2.5	7.8	13.0	84.5	1.5	5.2	8.4	91.3
Flickr30K	2.4	9.0	14.1	68.9	1.7	6.1	9.7	87.0
Flickr30K(+Entities)	4.5	13.6	21.7	48.5	2.6	7.6	11.9	85.1
Entities	3.1	12.1	19.4	53.9	2.2	7.0	10.9	85.1
Visual Genome	16.3	43.3	57.7	7.7	7.6	16.2	20.8	75.4

性能を表す．また，Med r は Median rank を指し，値が低いほど高い性能を表す．

実験結果を表3にまとめた．表3より，共通空間の性能向上には，**適切な粒度での訓練データ増加**が必要であることが読み取れる．例えば，訓練データの増加が性能向上に寄与した例として，Flickr8K から訓練したモデルと Flickr30K から訓練したモデルが挙げられる．Flickr30K での評価において，Flickr30K の方が Flickr8K よりも高い性能を示した．表2より，Flickr30K が Flickr8K の 4 倍の訓練データを含んでいることから，両者の性能の違いは，訓練データの量に起因すると考えられる．また，表3上，Visual Genome 上での評価における Flickr30K と Flickr30K(+Entities) の性能を比較すると，後者の方が高い性能を示した．Visual Genome の評価データは，物体画像と説明文から構成されるため，Flickr30K Entities 中に含まれる物体画像の訓練データが性能向上に貢献したと考えられる．

一方，適切な粒度でのデータ増加が実現されなかった例として，Visual Genome で訓練したモデルの Flickr30K での評価結果が挙げられる．ここで，Visual Genome の訓練データは Flickr8K の約 500 倍（表2）であるが，性能は Flickr8K で訓練したモデルよりも悪かった．これは，Flickr30K の評価データが画像全体と説明文から構成されており，物体画像のデータとは粒度が異なることが原因だと考えられる．

4.2 定性評価：共通空間の可視化

各データセットで訓練した結果を定性的に比較するため，得られた共通空間の可視化を行った．具体的には，2 つの訓練済みモデル（Flickr30K で訓練したものと Visual Genome で訓練したもの）と 2 種類の評価データ（Flickr30K と Visual Genome）を用いて，合計 4 種類の可視化を行った．

可視化した結果を図2に示す．理想的には，対応する画像と文の組は空間上で近傍に配置されることが望ましい．そのためには，図2a，2dのように，共通空間中で画像と文が混ざった形で配置されることが最低限必要である．しかし，図2b，2cに示した通り，訓練データと評価データの種類が異なる場合には，画像と文のクラスタが独立に形成された．粒度の異なるデータセットで訓練と評価を行った場合に，性能が著しく低かった（表3）のは，これが原因であると考えられる．

5 まとめ

本研究では，画像-言語のマルチモーダルタスクにおける再利用可能なコンポーネントの構築に向けて，画像と言語の情報の共通のベクトル空間への埋め込みを行った．画像全体と説明文，物体画像と説明文という粒度の異なる 2 つの埋め込み空間を構築し，両者の比較分析を行った．分析から，適切な粒度の訓練データを増やすことが重要であるとわかった．今後は，画像全体から物体画像と説明文を自動生成する（例: DenseCap[20]）ことで訓練データの拡張を行い，その効果を検証したい．

謝辞 本研究の一部は東北大学 Step-QI スクールおよび科研費（15H01702）の支援を受けて行った．

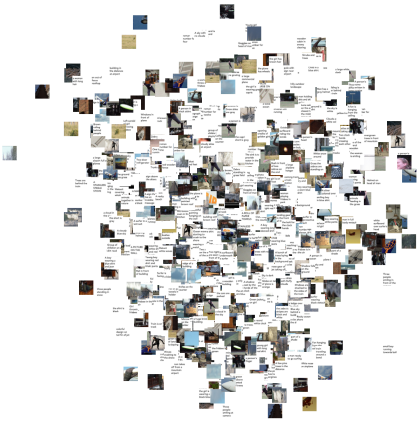
参考文献

[1] Oriol Vinyals et al. “Show and Tell: A neural image caption generator”. In: *CVPR*. 2015, pp. 3156–3164.

(a) Flickr30K で訓練後，Flickr30K を可視化したもの

(b) Visual Genome で訓練後，Flickr30K を可視化したもの

(c) Flickr30K で訓練後，Visual Genome を可視化したもの



(d) Visual Genome で訓練後，Visual Genome のデータを可視化したもの

図2: 各データセットで訓練した共通空間の可視化

[2] Jiuxiang Gu et al. “Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models”. In: *CVPR*. 2018, pp. 7181–7189.

[3] Jeffrey Pennington, Richard Socher, and Christopher Manning. “GloVe: Global Vectors for Word Representation”. In: *EMNLP*. 2014.

[4] Matthew Peters et al. “Deep contextualized word representations”. In: *NAACL*. 2018.

[5] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* (2018).

[6] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *ICLR*. 2015.

[7] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CVPR*. 2016, pp. 770–778.

[8] Satwik Kottur et al. “VisualWord2Vec (vis-w2v): Learning Visually Grounded Word Embeddings Using Abstract Scenes”. In: *CVPR*. 2016, pp. 4985–4994.

[9] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models”. In: *NIPS Deep Learning and Representation Learning Workshop*. 2014.

[10] Andrea Frome et al. “DeViSE: A Deep Visual-Semantic Embedding Model”. In: *NIPS*. 2013, pp. 2121–2129.

[11] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. “Deep Fragment Embeddings for Bidirectional Image Sentence Mapping”. In: *Advances in Neural Information Processing Systems 27*. 2014, pp. 1889–1897.

[12] Andrej Karpathy and Fei-Fei Li. “Deep Visual-Semantic Alignments for Generating Image Descriptions”. In: *CVPR*. 2015, pp. 3128–3137.

[13] Harsh Jhamtani and Taylor Berg-Kirkpatrick. “Learning to Describe Differences Between Pairs of Similar Images”. In: *EMNLP*. 2018, pp. 4024–4034.

[14] Micah Hodosh, Peter Young, and Julia Hockenmaier. “Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics”. In: *IJCAI*. 2015, pp. 4188–4192.

[15] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *TACL*. 2014.

[16] Cyrus Rashtchian et al. “Collecting Image Annotations Using Amazon’s Mechanical Turk”. In: *NAACL*. 2010.

[17] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *ECCV*. 2014.

[18] Ranjay Krishna et al. “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations”. In: *International Journal of Computer Vision*. 2017, pp. 32–73.

[19] Bryan A. Plummer et al. “Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models”. In: *International Journal of Computer Vision*. 2017, pp. 74–93.

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. “DenseCap: Fully Convolutional Localization Networks for Dense Captioning”. In: *CVPR*. 2016, pp. 4565–4574.